引用格式：员娇娇，胡永利，尹宝才．一种基于文本和图像的多模态目标检测方法[J]．中国传媒大学学报（自然科学版），2023，30(03)：41-49.
文章编号：1673-4793(2023)03-0041-09

一种基于文本和图像的多模态目标检测方法

员娇娇，胡永利，尹宝才*
（北京工业大学信息学部，北京 100124）

摘要：近年来，网络上涌现了大量的多模态数据（图像、文本、视频、音频等），由于不同模态的数据之间具有互补性，因此，利用不同模态的数据进行分类、检测、分割等任务已成为计算机视觉领域的研究热点。目标检测作为其中的一个重要方向，得到了越来越深入的研究。在传统的目标检测算法中，研究者们仅利用图像这一单模态的数据来实现对目标的分类和定位，这种做法没有考虑文本对目标检测算法性能的影响。本文重点研究基于文本和图像的多模态目标检测算法，首先利用传统的Faster R-CNN算法提取图像中的候选目标的特征，同时利用Bi-GRU算法提取文本的特征；其次，设计了一种有效的协同注意力模型来促进文本和图像这两种不同模态数据之间的融合。在大型的目标检测数据集MSCOCO上的实验结果表明，本文方法的检测精度高于仅利用图像信息的目标检测算法的精度，充分证明了本文方法的有效性。
关键词：多模态；目标检测；深度学习
中图分类号：TP37　**文献标识码：**A

A multimodal object detection method based on text and image

YUAN Jiaojiao, HU Yongli, YIN Baocai*
(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: In recent years, a large number of multimodal data (image, text, video, audio, etc.) have emerged on the network. Due to the complementarity between the multimodal data, it has become a research hotspot in the field of computer vision to use the data for the tasks of classification, detection, segmentation. As an important research direction in the field of computer vision, object detection has received more and more research. In the traditional object detection algorithm, researchers only use the single-mode data of the image to achieve the classification and location of the objects, which does not consider the impact of text on the performance of the object detection algorithms. This paper focuses on the object detection algorithm which based on text and images. Firstly, the traditional Faster R-CNN algorithm is used to extract the features of candidate objects in the image, and the Bi-GRU algorithm is used to extract the features of text; Secondly, an effective co-attention mode is designed to promote the interaction between text and images. The experimental results on MS COCO show that the detection accuracy of this method is higher than the object detection algorithm which only using image information, and the effective fusion of text and image is achieved.
Keywords: multimodal; object detection; deep learning

作者简介（*为通讯作者）：员娇娇（1990-），女，博士研究生，主要从事目标检测的研究。Email：yjj@emails.bjut.edu.cn；胡永利（1973-），男，博士生导师，教授，主要从事计算机视觉研究。Email：huyongli@bjut.edu.cn；尹宝才（1963-），男，博士生导师，教授，主要从事数字多媒体技术、多功能感知技术、虚拟现实与图形学方面的研究。Email：ybc@bjut.edu.cn

1 引言

随着互联网技术的发展,网络上产生了大量的多模态数据,例如图像、文本、视频、音频等[1]。在一些知名的社交媒体网站上(例如,微博、抖音、微信朋友圈、小红书、Flickr[2]、iTunes、YouTube等),使用者通常同时利用多种模态的数据来描述同一主题的事物,从而来分享他们的作品。以微信朋友圈为例,人们通常会在上传图像的同时,用文本来描述图像中的场景和事物。显然,这些不同模态的数据之间存在着天然的联系。近年来,随着大数据、深度学习[3]、人工智能[4]等技术的发展,对这些多模态数据之间的联系进行深度挖掘已成为商业开发和科学研究的热点,具有重要的现实意义和研究价值。诸如,在商业开发领域,基于文本的图像检索[5]在网站、淘宝等搜索功能中发挥着重要的作用;在科学研究领域也诞生了很多基于这些多模态数据的新兴研究方向,例如基于多模态数据的目标检测[6]、基于文本的图像指称分割[7]、基于多模态数据的文档分类[8]等。本文重点研究基于文本和图像这两种模态数据的目标检测。

目标检测[9]作为计算机视觉领域的研究热点,在很多工业领域和实际生活场景中发挥着重要的作用。按照网络框架的不同,目标检测算法可分为两阶段的目标检测算法和一阶段的目标检测算法[10]。其中,两阶段的目标检测算法,如Faster R-CNN[11]、Cascade R-CNN[12]等,其主要思想是先利用区域候选网络(Region Proposal Network, RPN)计算得到图像中的目标候选框(包含大量的前景区域和少量的背景区域),然后再对这些候选框的类别和坐标进行拟合;一阶段的目标检测算法,如YOLO[13]、SSD[14]、RefineDet[15]等,其不需要区分前景区域和背景区域,而是利用预先设置好的锚框对图像中的目标进行直接预测。无论是一阶段的目标检测算法还是两阶段的目标检测算法,在传统的做法中,只利用图像这一个模态的数据进行计算,这种做法需要收集大量的图像并对图像中物体的类别和坐标进行人工标注,数据的收集成本较高,目标检测算法的性能很大程度上依赖于图像的数量和人工标注的准确度。尽管目前这种只基于图像的目标检测算法取得了良好的性能,但是随着多模态数据的出现以及基于多模态数据的研究工作的深入,这种传统的做法忽略了其他模态的数据中包含的丰富信息,造成了信息的浪费。

因此,受相关的多模态研究工作的启发,我们观察到网络上有大量的图像、文本对数据,并且与之相关的研究工作也取得了巨大的进展,极大的促进了分类、检测、分割、图像检索、视觉问答[16]等领域的发展。例如,基于文本的图像检索算法是实现网站搜索引擎的重要技术,在学术界也得到了广泛的关注和研究。Ge等[17]提出了一种跨模态的语义增强交互模型用于图像和文本之间的检索。该模型首先设计了一种空间和语义关系图来实现图像内目标之间的关系推理,从而来增强图像的特征表达;然后分别通过将目标特征和句子特征、单词特征和图像特征进行交互来增强视觉和文本的特征。该模型在Flicker数据集上的rSum指标达到了521.3,将检索性能提升了8.1%。Mengjun等人[18]提出了一种ViSTA算法,该算法通过将图像分解成指定大小的图像块,并和句子中的单词一起送入Transformer[19]模块中进行特征融合从而来提升跨模态检索的性能。相比于其他的跨模态检索算法,ViSTA在Recall@1指标上相对提升了8.4%。在多模态图像分类领域,Yuan等人[20]提出了一种文本辅助的图像分类算法,该算法利用词袋模型[21]提取文本的特征,并将文本特征映射到对应的图像特征空间,然后将这两种特征进行融合。与传统的在ImageNet等[22]大型图像数据集上训练得到的图像分类算法相比,该算法在仅利用少量标注的图像和大量文本的情况下取得了近似的图像分类性能,该算法充分利用文本的信息来减少分类算法对图像标注的依赖。在图像分割领域,Li等人[23]提出了一种文本驱动的语义分割模型,通过将图像的语义标签转化为特定维度的文本特征来控制分割输出的类别和类别数量,在零样本的图像语义分割任务上,算法性能提升了5%。特别的,在多模态目标检测领域,Aishwarya Kamath等人[24]在DETR[25]模型的基础上提出了一种MDETR算法。该算法将图像和文本都作为模型的输入,利用卷积神经网络(Convolutional Neural Network, CNN)[26]提取图像的特征,利用RoBERTa[27]提取文本的特征,然后对这两种特征进行拼接,将融合之后的特征送入到基于Transformer构建的检测模块中进行检测,极大地提升了目标检测的平均精度(Average Precision, AP)。值得一提的是,针对目标检测中类别不平衡的问题(一些不常见的目标类别由于较少的训练样本而造成检测精度较低),在MDETR中由于文本信息的加入,其检测精度得到了显著的提升(4%左右),这充分说明了文本中的语义信息能够有效提升图像的特征表达。为了减少对人工标注数据的依赖,

Liu等人[28]利用文本和鼠标的轨迹作为监督信息，并设计了相应的对比损失来实现文本和图像在特征空间的对齐，从而从文本中获取丰富的语义信息来得到有效的视觉特征表示。在MSCOCO数据集[29]上的结果显示，该模型在取得近似的目标分类和定位的性能下，仅利用了10%的数据量，显示了文本对于图像中目标分类和定位的巨大价值。Zhong等人[30]提出了RegionCLIP模型，该模型利用CLIP算法[31]来对齐图像中目标的视觉特征和文本中的单词特征，从而得到一个基于文本和图像的预训练模型，该预训练模型能有效促进基于文本和图像的多模态目标检测的性能。将RegionCLIP应用到开放世界的目标检测任务中时，在COCO数据集上，新类别的检测性能得到了2.2% -3.8%左右的提升，这也再次验证了文本对于图像目标检测的重要性，充分说明了不同模态数据之间的互补性。

基于上述研究，为了进一步充分挖掘文本中的信息对于图像目标检测的作用，实现不同模态特征之间的深度融合，本文在Faster R-CNN的基础上提出了一种基于文本和图像的多模态目标检测方法。首先，利用Faster R-CNN算法提取图像中目标的特征，同时利用Bi-GRU[32]算法提取文本中每个单词的特征；为了实现这两种特征的深入融合，本文还设计了一种有效的协同注意力模型来对齐文本特征和图像中目标区域的特征，将对齐后的文本特征融合到对应的视觉特征中来增强视觉特征的判别性，从而提升目标检测的性能。为了验证本文模型的有效性，我们在MSCOCO数据集上进行了实验，结果显示本文方法能有效利用文本信息来提升图像目标检测的性能。

2 本文算法

为了详细阐述本文提出的基于文本和图像的多模态目标检测方法，该部分首先介绍模型的整体框架，然后介绍其中的主要功能模块。

2.1 网络结构

本文模型的整体框架如图1所示。该模型以Faster R-CNN为基础，将成对的图像和文本作为模型的输入，其中，图像首先经过Faster R-CNN的主干网络提取整幅图像的特征，然后将该特征输入到RPN网络中得到图像中前景的特征向量；与此同时，将文本输入Bi-GRU中提取每个单词的特征向量。接下来，为了充分利用文本特征辅助图像特征，将提取的图像的特征向量和文本的特征向量输入到协同注意力模块中，将两种不同模态的特征进行融合，从而得到被文本特征增强之后的图像特征。最后，用增强之后的特征输入到Faster R-CNN后续的检测模块(即目标分类模块和目标定位模块)中进行目标检测。

2.2 图像特征提取模块

在Faster R-CNN算法中，RPN网络输出特定数目的目标候选区域(包含大量的前景区域和少量的背景区域)，将这些区域的坐标映射到整幅图像的特征图后得到这些区域对应的特征向量，从而得到整幅图像的特征表示矩阵H^v，其中$H^v=\{h_1,\cdots,h_m\}\in\mathbb{R}^{m\times d_1}$，$m$指的是目标候选区域的个数，$d_1$指的是每个目标候选区域对应的图像特征向量的维度。

2.3 文本特征提取模块

研究表明，在利用文本信息(通常以句子的形式出现)进行多模态目标检测的过程中，文本中的单词会强调图像中的相关区域，从而有利于目标检测性能的提升。由于文本中的单词存在较强的上下文关联性，因此，本文使用Bi-GRU算法作为文本的特征编码器，从而得到每个单词的特征。

利用Bi-GRU算法来提取单词特征的公式如下：

$$\begin{aligned}\overrightarrow{h_i}&=\overrightarrow{GRU}(x_i),i\in[1,n]\\\overleftarrow{h_i}&=\overleftarrow{GRU}(x_i),i\in[1,n]\\e_i&=\frac{(\overrightarrow{h_i}+\overleftarrow{h_i})}{2}\end{aligned}\tag{1}$$

其中，x_i指的是句子中的每个单词，n指的是句子中单词的个数，$\overrightarrow{h_i}$和$\overleftarrow{h_i}$分别指的是按照从前到后和从后到前的顺序将句子中的每个单词输入GRU[32]算法中得到的单词的特征向量，e_i指的是最终每个单词的特征向量，维度为d_2。经过Bi-GRU算法得到整个句子的特征矩阵$H^x=\{e_1,\cdots,e_n\}\in\mathbb{R}^{n\times d_2}$。

2.4 协同注意力模块

在得到图像的特征矩阵H^v和文本的特征矩阵H^x之后，为了实现两种模态信息之间的交互，充分挖掘文本特征对图像特征的互补能力，我们提出了一种协同注意力模块。该模块的计算过程如图2所示，其主要思想是：首先，通过计算H^v中的每个特征向量对H^x中所有特征向量的注意力权重A^x，从而得到每个图像特征和所有文本特征之间的相似度，相似度越大，则

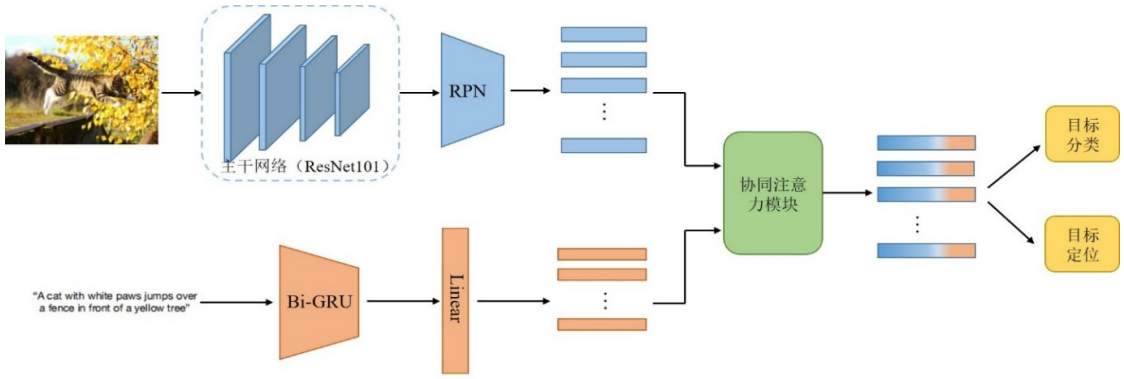

图1 本文模型的整体框架

表示该文本特征和图像特征之间的语义信息越接近。然后,将该相似度作为权重对相应的文本特征进行加权求和,得到新的特征矩阵$\hat{H}^x$。将$\hat{H}^x$和H^v进行拼接,从而将文本特征融入到图像特征中得到增强之后的图像特征$\hat{H}^v$,实现文本特征对图像特征的增强。

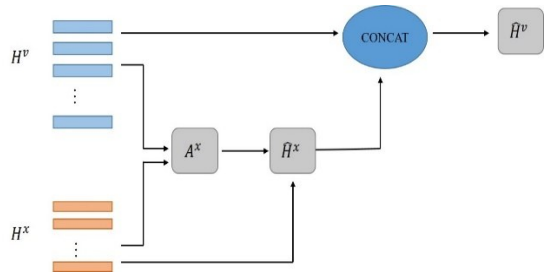

图2 协同注意力模块

如图2所示,本文提出的协同注意力模块的具体计算过程如下:

步骤1:计算注意力权重矩阵A^x。

A^x的计算过程如公式(2)和公式(3)所示:

$$S = H^v W (H^x)^T \tag{2}$$

$$A^x = soft\max(S) \in \mathbb{R}^{m \times n} \tag{3}$$

其中,$S \in \mathbb{R}^{m \times n}$是$H^v$和$H^x$之间的相似度矩阵,$W \in \mathbb{R}^{d_1 \times d_2}$是通过神经网络训练得到的,$S_{ij}$表示第$i$个图像特征向量和第$j$个文本特征向量之间的相似度。对$S$按行进行$soft\max$操作得到图像特征对文本特征的注意力权重矩阵$A^x$。其中,$A^x_{ij}$表示第$i$个图像特征向量对第$j$个文本特征向量的注意力值。

步骤2:计算特征矩阵$\hat{H}^x$。

将A^x和H^x进行矩阵相乘得到新的特征矩阵$\hat{H}^x$,$\hat{H}^x$的计算过程如公式(4)所示。

$$\hat{H}^x = A^x H^x \tag{4}$$

其中,$\hat{H}^x \in \mathbb{R}^{m \times d_2}$。

步骤3:计算被文本特征增强的图像特征$\hat{H}^v$。$\hat{H}^v$的计算过程如公式(5)所示。

$$\hat{H}^v = Concat(H^v, \hat{H}^x) \tag{5}$$

其中,$\hat{H}^v \in \mathbb{R}^{m \times (d_1 + d_2)}$,$\hat{H}^v$共有$m$个特征向量,每个特征向量都包含了图像和文本两种模态的信息。

通过上述步骤,利用提出的协同注意力模块得到了被文本特征增强的图像特征$\hat{H}^v$,最终将该特征送入Faster R-CNN后续的检测模块中对图像中的物体进行检测。

3 实验结果与分析

3.1 数据集

为了充分验证本文模型的有效性,我们在MSCOCO数据集上进行了相关实验。MSCOCO数据集是一个被用来研究多种计算机视觉任务(目标检测、语义分割、视觉问答等)的大型数据集。它有超过330K张图像,每张图像都包含对应的5句文本描述,这些目标的类别和坐标都进行了详细的标注。在本文的实验中,我们分别在MSCOCO2017数据集和MSCOCO2014数据集上进行了实验。MSCOCO2017数据集的具体划分如下:训练集共包含118287张图像,验证集包含5000张图像。MSCOCO2014数据集的具体划分如下:训练集共包含82783张图像,验证集包含40504张图像。无论是训练集还是验证集,每张图像都包含相应的5句文本描述。在MSCOCO数据集上的实验均在这两个数据集的训练集上进行训练,验证集上进行测试。

此外,为了进一步验证算法的性能,本文还利用了Visual Genome[33]数据集来验证本文方法的有效性。该数据集是一个被广泛使用的数据集,每张图像除了包含相应的目标框以外,还包含其对应的文本表述以

及这些目标框之间的关系。该数据集共包含108077张图像，其中训练集包括75651张图像，测试集包括32422张图像。

3.2 评价指标及参数设置

（1）评价指标

按照MSCOCO数据集的方式，为了更加直观地分析文本信息对目标检测算法的贡献，本文采用的评价指标如表1所示。这些评价指标的数值越高，则算法的性能越好。其中，交并比（Intersection Over Union, IOU）是目标检测中使用的一个概念，IOU的计算通过用检测框和真值框的交集除以它们的并集得到。

表1 评价指标

评价指标	释义
AP^{50}	IOU=0.5时的平均精度
AP^{75}	IOU=0.75时的平均精度
AP^{s}	对小尺寸目标的平均精度
AP^{m}	对中等尺寸目标的平均精度
AP^{l}	对大尺寸目标的平均精度
mAP	对所有类别的平均精度取均值

（2）实验环境及参数设置

在我们的实验中，在图像特征提取部分，采用ResNet101[34]作为Faster R-CNN的主干网络，和RPN网络一起来提取所需的特征向量；在文本特征提取部分，随机从每张图像对应的5句文本描述中选择一句作为本文模型所需的文本。由于文本的长度是不一样的，为了便于计算，将文本的长度设置为15（即文本最多包含15个单词）。在实验的过程中，若文本的长度大于15，则进行截断，若长度小于15，则将空缺的部分填充为空字符。经过这样的处理，输入端的文本长度保持一致，便于计算。其中，$m = 128$，$n = 15$，$d_1 = 512$，$d_2 = 512$。

此外，在此基础上，考虑到Cascade R-CNN算法也是典型的两阶段目标检测算法，其将多个弱检测器进行级联从而形成强的目标检测器。在Cascade R-CNN算法中，每个阶段检测器的输入候选框是前一阶段检测器的输出候选框。所以在本文的实验中，也进一步基于Cascade R-CNN算法进行了相关的实验。其中，在每个阶段的检测器中，文本和图像信息都以图1中的方式进行信息交互，即在Cascade R-CNN算法中有几个阶段的检测器，则该交互过程重复几次。同时，由于Mask R-CNN[35]中也采用RPN网络来生成候选区域，本文也将其作为一个基线在各个数据集上进行实验。

本文模型基于NVIDIA GeForce RTX 3090 GPUs平台进行开发，采用Pytorch框架来构建网络模型。为了使模型得到充分的训练，本文采用随机梯度下降（Stochastic Gradient Descent，SGD）优化器，将学习率设置为0.001，在每一次迭代中输入15个文本-图像对，迭代次数设置为300k。

3.3 实验结果

目前，由于基于文本和图像的多模态目标检测工作较少，我们将本实验中用到的各个模型标记如下：

V1：传统的Faster R-CNN算法，即只有图像作为输入；

V2：多模态的Faster R-CNN算法，即本文中方法，将图像和文本同时作为输入；

V3：传统的Cascade R-CNN算法，即只有图像作为输入；

V4：多模态的Cascade R-CNN算法，即在每个阶段的检测器中都将文本和图像的信息按照图1中的方式进行交互。

V5：传统的Mask R-CNN算法，即只有图像作为输入；

V6：多模态的Mask R-CNN算法，即按照本文的方法将图像和文本同时作为输入，并在RPN模块部分进行交互。

下面分别介绍上述六个模型在各个数据集上的检测结果。

（1）在MSCOCO2014和MSCOCO2017上的结果

为了证明本文模型的有效性，分别在MSCOCO2014和MSCOCO2017数据集上开展了实验。其中，在MSCOCO2014上的实验结果如表2所示。

从表2可以看出，本文方法能够有效利用文本信息提升图像目标检测的性能。其中，在Faster R-CNN算法中，对于大尺度目标，由于文本信息的加入，其检测性能得到了较大程度上的提升（AP_l从43.6提升到45.0）；对于中等尺度的目标，其检测性能也得到了一定程度的提升（AP_m从31.9提升到32.3）；对于小尺度的目标，其性能基本上没有变化，这是因为本文仅利用了主干网络的最后一层特征来提取图像的特征向量，而该层中含有较多的大尺度目标和中等尺度目标的信息，小尺度目标的信息则较少，所以对小尺度目标的检测性能的提升不是很明显。

如表2所示，在Cascade R-CNN算法中，当在每个

表2 各模型在MSCOCO2014数据集上的性能比较

方法	评价指标					
	AP^{50}	AP^{75}	AP_s	AP_m	AP_l	mAP
V1	48.5	29.4	11.2	31.9	43.6	28.3
V2	47.5	30.7	11.0	32.3	45.0	28.9
V3	62.1	46.3	23.7	45.5	55.2	42.7
V4	64.5	47.3	25.6	47.7	57.0	44.6
V5	60.3	41.7	20.1	41.1	50.2	38.2
V6	61.4	42.0	21.8	42.3	51.5	40.6

阶段的检测器中加入文本信息进行交互时,其各项检测精度的指标得到了明显的提升,小目标的检测性能提升了1.9,中等尺度目标的检测性能提升了2.2,大尺度目标的检测性能提升了1.8。相比于Faster R-CNN算法,本文的方法在Cascade R-CNN算法上的提升效果更明显,这得益于Cascade R-CNN算法的级联结构,在这种结构下文本和图像也相应地进行了多阶段的交互。以上结果充分说明了本文方法的合理性,同时也表明了对于存在RPN结构的检测模型,本文的方法可以实现即插即用的效果,模型结构具有合理性和易用性。同样的,在Mask R-CNN算法中也验证了本文方法的有效性。

为了在多个不同的数据集上验证本文的方法,我们也在MSCOCO2017数据集上进行了实验和对比。其结果如表3所示,从结果中可以看出,在不同的两阶段目标检测算法中,加入文本信息之后都能有效提升检测器的性能。

为了更加详细的分析文本对于图像目标检测的作用,我们将V1和V2在MSCOCO2017中的图像上的检测结果进行了可视化的展示,结果如图3所示。图3的左边是V1的检测结果,右边是V2的检测结果。通过对比我们发现,在Faster R-CNN算法中,一些物体没有被检测出来(例如,第一张图片中椅子旁边的鸟,第二张图片中的门,第三张图片中间的人等)。而在本文方法中由于文本信息的加入能够有效地将这些物体检测出来,这归功于文本中丰富的语义信息对图像语义的增强,充分说明了不同模态的数据之间具有互补性。当将文本和图像数据以有效的方式进行交互时,充分利用不同模态的数据之间的互补性,可有效提高图像中目标的检测性能。

(2)在Visual Genome数据集上的检测结果

Visual Genome也是一个被广泛用于进行多模态视觉任务研究的数据集,其中的图像也包含了相应的目标框标注及自然语言描述,在视觉问答、场景图的构建等任务中发挥着重要的作用。值得注意的是,和MSCOCO数据集相比,Visual Genome中目标的尺度更加丰富多变,且数据集在标注的过程中存在较为严重的偏置,即一些常见的目标被大量标注,而部分目标的标注则较少,这就导致了不同目标的训练样本在数量上存在较大的差异;此外,该数据集在标注的过程中很多类别在语义上是重叠的(比如person和man),一些类别在标注的过程中也存在单数和复数的区别(比如person和people),还有一些类别无法被精确定位(比如街道和田野)。以上这些因素都造成该数据集对目标检测、视觉问答等任务具有较大的挑战。为了进一步验证本文方法的有效性,本文在Visual Genome数据集上也开展了多模态目标检测任务的实验,以验证在上述挑战中文本对图像目标检测的作用。各个目标检测器在该数据集上的实验结果如表4所示。

从表4中可以看出,相比于MSCOCO数据集,Visual Genome数据集还是给各个检测器带来了一定的挑战。但是本文的方法依旧能够稳定提升相应的目标检测性能(即V2,V4,V6的性能分别高于V1,V3,V5),再次验证了跨模态的信息交互对于目标检测性能的提升作用。

为了进一步直观的展示本文方法的有效性,我们将V1和V2在Visual Genome数据集上的检测结果进行可视化,以进行进一步的比较分析,结果如图4所示。其中左边是V1的检测结果,右边是V2的检测结果。从图中可以看出,当仅使用图像信息时,第一张图像中的“trash can”和第二张图像中的“hand”没有被检测出来;而当加入文本信息时,漏检的物体则成功被检测到。

图3 在MSCOCO2017数据集上的可视化分析

表3 各模型在MSCOCO2017数据集上的性能比较

方法	评价指标					
	AP^{50}	AP^{75}	AP_s	AP_m	AP_l	mAP
V1	50.4	31.1	14.6	35.2	45.6	31.9
V2	50.9	32.3	14.8	36.7	46.9	32.8
V3	63.2	47.5	25.0	46.9	57.1	43.9
V4	66.0	48.1	27.0	49.6	59.0	47.3
V5	61.8	42.2	22.5	42.7	51.9	39.0
V6	62.7	43.5	22.7	43.6	52.8	41.8

表4 各模型在Visual Genome数据集上的性能比较

方法	评价指标					
	AP^{50}	AP^{75}	AP_s	AP_m	AP_l	mAP
V1	21.7	30.8	9.3	10.6	11.5	10.5
V2	22.6	33.7	11.5	12.3	13.2	12.3
V3	24.7	34.6	12.3	14.5	15.2	14.0
V4	25.6	35.9	13.1	15.8	15.7	15.6
V5	22.4	30.6	14.5	16.7	15.0	13.7
V6	23.4	31.3	14.6	16.8	16.3	13.9

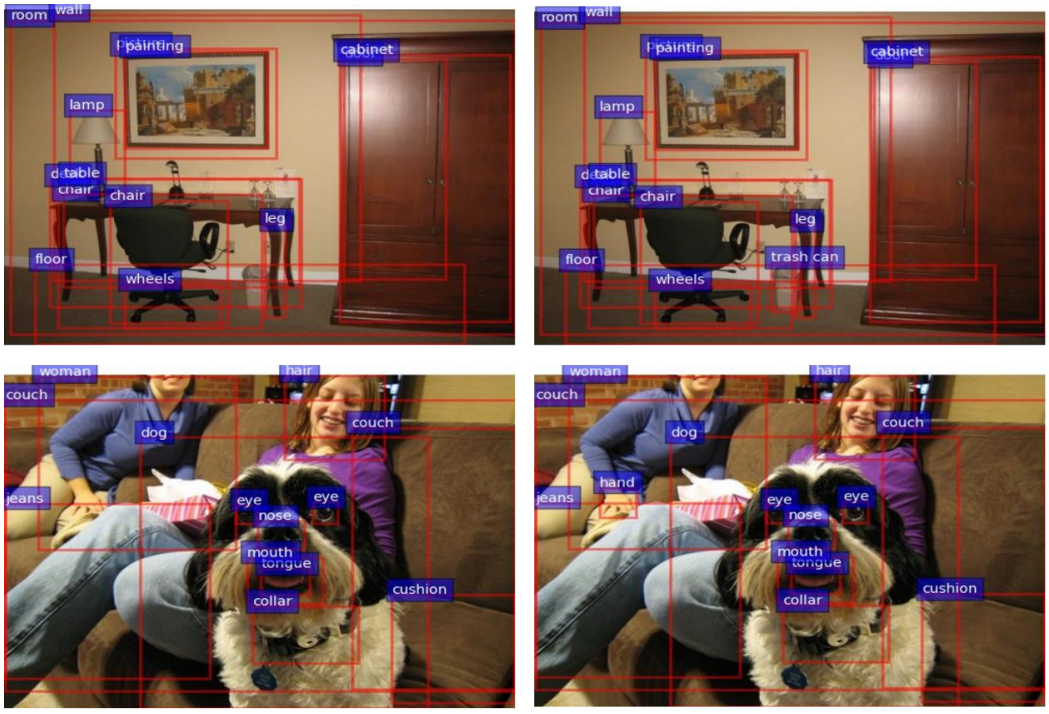

图4 在Visual Genome数据集上的可视化分析

通过将本文提出的文本和图像特征交互模块在多个不同的检测器以及数据集上进行实验,充分验证了利用不同模态的信息之间的互补性可有利于提升传统的检测器的性能。

4 结论

本文提出了一种基于文本和图像的多模态目标检测方法。为了探索文本数据对图像目标检测的作用,在传统的Faster R-CNN算法的基础上,基于设计的协同注意力模型将文本特征和图像特征进行深度融合,从而利用文本特征来对图像特征进行增强。在此基础上,在其他的两阶段的检测器中也验证了本文方法的有效性。实验结果表明,该方法能够有效利用文本信息来提高图像中不同尺度目标的检测性能,探索了多模态数据对于目标检测的优势。在未来的工作中,我们还将探索将文本和图像在不同层次的特征上进行多层交互,从而来进一步提升图像中不同尺度目标的检测性能。

参考文献(References):

[1] 李玉腾, 史操, 许灿辉, 程远志. 基于视觉和文本的多模态文档图像目标检测[J]. 计算机应用研究, 2023, 40(4).

[2] Huiskes M, Lew M. The MIR flickr retrieval evaluation[C]// ACM International Conference on Multimedia Information Retrieval, 2008: 39-43.

[3] 孙志军, 薛磊, 许阳明, 王正. 深度学习研究综述[J]. 计算机应用研究, 2012, 29(8): 5.

[4] 夏定纯, 徐涛. 人工智能技术与方法[M]. 武汉:华中科技大学出版社, 2004.

[5] Li J, Xu X, Yu W, et al. Hybrid fusion with intra- and cross-modality attention for image-recipe retrieval[C]//SIGIR '21: The 44th International ACM SIGIR Conference on Research

and Development in Information Retrieval, 2021: 244-254.
[6] Chen Z, Lin Q, Sun J, et al. Cascaded cross-modality fusion network for 3D object detection[J]. Sensors, 2020, 20(24): 7243.
[7] Zhou Q, Hui T, Wang R, et al. Attentive excitation and aggregation for bilingual referring image segmentation[J]. ACM Transactions on Intelligent Systems and Technology, 2021, 12(2): 1-17.
[8] Ko Y, Seo J. Issues and empirical results for improving text classification [J]. Journal of Computing Science & Engineering, 2011, 5(2): 150-160.
[9] Wang G, Ding H, Li B, et al. Trident-YOLO: improving the precision and speed of mobile device object detection[J]. IET Image Processing, 2022(1): 16.
[10] 王婉婷, 姜国龙, 褚云飞, 陈业红. 从RCNN到YOLO系列的物体检测系统综述[J]. 齐鲁工业大学学报, 2021, 35(5): 8.
[11] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.
[12] Wu Y, Liu W, Wan S. Multiple attention encoded cascade R-CNN for scene text detection [J]. Journal of Visual Communication and Image Representation, 2021(11): 103261.
[13] Redmon J, Farhadi A. YOLO9000: better, faster, stronger [C]//IEEE Conference on Computer Vision & Pattern Recognition, 2017: 6517-6525.
[14] Liu W, Anguelov D, Erhan D, et al. Ssd:single shot multibox detector [C]//European Conference on Computer Vision, 2016: 21-37.
[15] Zhang S, Wen L, Bian X, et al. Single-shot refinement neural network for object detection [C]//IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4203-4212.
[16] Pan Y, Li Z, Zhang L, et al. Causal inference with knowledge distilling and curriculum learning for unbiased VQA[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2022, 18(3): 67: 1-67: 23.
[17] Ge X, Chen F, Xu S, et al. Cross-modal semantic enhanced interaction for image-sentence retrieval[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023: 1022-1031.
[18] Cheng M, Sun Y, Wang L, et al. ViSTA: vision and scene text aggregation for cross-modal retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5184-5193.
[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Annual Conference on Neural Information Processing Systems, 2017: 5998 - 6008.
[20] Lin Y, Chen Y, Xue G R, et al. Text-aided image classification: using labeled text from web to help image classification[C]//International Asia-pacific Web Conference, 2010: 267-273.
[21] Silva F B, Werneck R, Goldenstein S K, et al. Graph-based bag-of-words for classification [J]. Pattern Recognition, 2018, 74: 266-285.
[22] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25(2).
[23] Li B, Weinberger K Q, Belongie S, et al. Language-driven semantic segmentation[C]//The Tenth International Conference on Learning Representations, 2022.
[24] Kamath A, Singh M, LeCun Y, et al. MDETR - modulated detection for end-to-end multi-modal understanding[C]//IEEE International Conference on Computer Vision, 2021: 1760-1770.
[25] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers [C]//European Conference on Computer Vision, 2020: 213-229.
[26] Chua L O, Roska T. The CNN paradigm[J]. IEEE Transactions on Circuits & Systems I: Fundamental Theory & Applications, 1993, 40(3): 147-156.
[27] Liu Y, Ott M, Goyal N, et al. Roberta: a robustly optimized bert pretraining approach[J]. CoRR, 2019.
[28] Liu Z, Stent S, Li J, et al. LocTex: learning data-efficient visual representations from localized textual supervision[C]// IEEE International Conference on Computer Vision, 2021: 2147-2156.
[29] Lin T Y, Maire M, Belongie S J, et al. Microsoft coco: common objects in context[C]// European Conference on Computer Vision, 2014: 740 - 755.
[30] Zhong Y, Yang J, Zhang P, et al. RegionCLIP: region–based language-image pretraining[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2022: 16772-16782.
[31] Radford A, Kim J, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the 38th International Conference on Machine Learning, 2021: 8748-8763.
[32] Wu C. Text sentiment classification based on BERT embedding and sliced multi-head self-attention Bi-GRU[J]. Sensors, 2023, 23.
[33] Krishna R, Zhu Y, Groth O, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123: 32-73.
[34] 李爱莲, 刘浩楠, 郭志斌, 解韶峰, 崔桂梅. 改进ResNet101网络下渣出钢状态识别研究[J]. 中国测试, 2020, 46(11): 5.
[35] He K, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961-2969.

编辑:赵志军,龙学锋